Extraction of SNOMED Concepts from Medical Record Texts

Diane E. Oliver, MD and Russ B. Altman, MD, PhD
Section on Medical Informatics
Stanford University School of Medicine
Stanford, CA 94305
oliver@camis.stanford.edu, altman@camis.stanford.edu

ABSTRACT

Clinicians have traditionally documented patient data using natural language text. With the increasing prevalence of computer systems in health care, an increasing amount of medical record text will be stored electronically. However, for such textual documents to be indexed, shared, and processed adequately by computers, it will be important to be able to identify concepts in the documents using a common medical terminology. Automated methods for extracting concepts in a standard terminology would enhance retrieval and analysis of medical record data. This paper discusses a method for extracting concepts from medical record documents using the medical terminology SNOMED-III (Systematized Nomenclature of Human and Veterinary Medicine, Version III). The technique employs a linear least squares fit that maps training set phrases to SNOMED concepts. This mapping can be used for unknown text inputs in the same domain as the training set to predict SNOMED concepts that are contained in the document. We have implemented the method in the domain of congestive heart failure for history and physical exam texts. Our system has a reasonable response time. We tested the system over a range of thresholds. The system performed with 90% sensitivity and 83% specificity at the lowest threshold, and 42% sensitivity and 99.9% specificity at the highest threshold.

INTRODUCTION

Although computers show much promise for improving storage and access of medical records, retrieval and analysis will be difficult if records are stored as natural language text with inadequate indexing and processing to establish semantic content. Major barriers in the effort toward the development of a computer-based patient record in medical care are the lack of a standardized medical terminology and the ability to code medical record text using such a standard. Clinicians do use a constrained vocabulary in their patient records, but like natural language, there is still a fair amount of variability in how they might express ideas in writing. Clinicians cannot be expected to learn a standardized vocabulary and will either require guidance for data entry through a structured computer interface, or text written by clinicians will have to be analyzed automatically for concepts in a common vocabulary.

In this paper, we propose a method for automated mapping of natural language text in medical record documents to concepts in a controlled medical terminology. The method is based on using a training set of "free text to terminology" mappings using a linear least squares fit (LLSF) and singular value decomposition (SVD) technique. The medical terminology selected for this task is SNOMED-III (Systematized Nomenclature of Human and Veterinary Medicine) [1], although our method is relevant to any terminology. SNOMED is a medical terminology developed by the College of American Pathology that aims to cover broad areas in clinical medicine. It includes terms for anatomy, morphology, signs and symptoms, living organisms, drugs, occupations, devices and activities associated with disease, social context, diagnoses, procedures, and modifiers.

There are a number of potential uses for automated extraction of controlled terminology concepts from medical record text documents. First, if patient data are stored electronically in text format and the volume of data is large for a given patient, searching through a patient's record can be simplified if the data is indexed using a controlled terminology. Second, if queries are to be made across a population of patients in a database made up of textual documents, retrieval can be enhanced if the content of the texts is based on a standardized terminology. Queries across patient populations and often across different databases are important for research on practice patterns, retrospective clinical studies, linking of costs to processes of care, and identification of patients who are eligible to be included in clinical trial protocols. Third, decision support systems that are based on the common controlled terminology could trigger alerts or recommendations on patients whose textual medical records contained the relevant content. Finally, this method could be used to compare the value of one

0195-4210/94/$5.00

terminology to another by looking at the performance of each terminology when applied to the same set of data.

Yang and Chute [2, 3] used a linear least squares fit approach to map physician-recorded diagnoses to ICD-9-CM codes [4]. Our method follows a similar approach, but rather than trying to identify the single most likely code for an input text, we map an input text to a group of relevant SNOMED codes that collectively represent the content of the input text. Yang and Chute found the LLSF method to be superior to string matching, statistical weighting, and latent semantic indexing in their application domain.

METHODS

The methodology described here and implemented in our system is an LLSF approach. The mapping learns a linear function that maps texts to SNOMED concepts on a training set of data. It then can predict SNOMED terms for unknown input texts.

Data were collected for the training set from a set of 21 medical record hospital admission summaries of patients diagnosed with congestive heart failure. Phrases that were deemed of medical importance for patients with congestive heart failure and that could be represented in SNOMED were selected from these documents by a person who has experience in the practice of medicine as well as in the use of SNOMED. Phrases were selected from the chief complaint, history of present illness, and physical examination sections of the medical record. In the physical examination, only lung, heart, and extremity examinations were included.

Each text phrase was matched with one or more SNOMED concepts. The training set data, which consists of the text phrases and their corresponding SNOMED concepts, is stored in two matrices: Matrix **A** contains data on the words in the training set text phrases, and matrix **B** contains data on the SNOMED terms selected as equivalent in meaning to concepts in the training set phrases. A total of 197 training set phrases were included. The number of distinct words in the union of all the words in the training set is 365. The number of distinct SNOMED terms in the union of all the SNOMED terms selected for the training set is 139. Characteristics of matrices **A** and **B** are described below.

Matrix **A**

(1) There are 197 columns with one column for each text phrase.
(2) There are 365 rows with one row for each word that is found one or more times in the training set.
(3) An entry $\mathbf{A}_{ij}$ is set to 1 if word i is found in phrase j.
(4) An entry $\mathbf{A}_{ij}$ is set to 0 if word i is not found in phrase j.

Matrix **B**

(1) There are 197 columns with one column for each text phrase.
(2) There are 139 rows with one row for each SNOMED term that is found one or more times in the training set.
(3) An entry $\mathbf{B}_{ij}$ is set to 1 if SNOMED term i is relevant to phrase j.
(4) An entry $\mathbf{B}_{ij}$ is set to 0 if SNOMED term i is not relevant to phrase j.

Using a linear least squares fit for the data, a mapping matrix **W** was calculated that optimally solves the equation $\mathbf{WA} = \mathbf{B}$. Then for an unknown input vector, **a**,

$$\mathbf{Wa} = \mathbf{b} \qquad (1)$$

where the predicted output values are in vector **b**.

Since solving for **W** does not always yield an exact solution, the goal is to find an appropriate **W** that minimizes the error in $\mathbf{WA} - \mathbf{B}$. A measure of this error is the sum of the squares of the entries in matrix $\mathbf{E} = \mathbf{WA} - \mathbf{B}$. That is, if $\mathbf{E} = \mathbf{WA} - \mathbf{B}$ is an $m \times k$ matrix, then the value to minimize is

$$\sum_{i=1}^{k} \sum_{j=1}^{m} \mathbf{E}_{ij}^2 \qquad (2)$$

where $\mathbf{E}_{ij}$ is the *ith* row and *jth* column of matrix **E**.

A commonly used method for solving a linear least squares fit problem is based on a matrix factorization technique known as singular value decomposition (SVD) [5]. For an $n \times k$ matrix **A** and an $m \times k$ matrix **B**, the computation for an LLSF for $\mathbf{WA} = \mathbf{B}$ is as follows:

Compute the SVD of matrix **A**. That is, determine **U**, **S**, and **V** such that $\mathbf{A} = \mathbf{USV}^{\mathbf{T}}$. **A** is the $n \times k$ matrix being decomposed, **U** is an $n \times p$ orthogonal matrix, S is a $p \times p$ diagonal matrix with all positive values on the diagonal, and $\mathbf{V}^{\mathbf{T}}$ is a $p \times k$ orthogonal matrix where $\mathbf{V}^{\mathbf{T}}$ is the transpose of **V**. Since **U** and $\mathbf{V}^{\mathbf{T}}$ are orthogonal, they can be multiplied by their transposes to yield the identity matrix. This fact is used in the following sequence of matrix manipulations to find an equation for **W**.

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}} \quad (3)$$

$$\mathbf{W}\mathbf{A} = \mathbf{B} \quad (4)$$

$$\mathbf{W}\mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}} = \mathbf{B} \quad (5)$$

$$\mathbf{W} = \mathbf{B}\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\mathbf{T}} \quad (6)$$

where $\mathbf{S}^{-1}$ is the inverse of **S** and $\mathbf{U}^{\mathbf{T}}$ is the transpose of **U**.

Therefore, the SVD approach allows us to calculate a matrix **W** that solves the linear least squares fit problem **WA** = **B**. The resulting matrix **W** is $m \times n$. In the training set, there are k text phrases, n distinct words, and m distinct SNOMED terms.

$$\underset{m \times n}{\mathbf{W}} \times \underset{n \times k}{\mathbf{A}} = \underset{m \times k}{\mathbf{B}} \quad (7)$$

We used a published algorithm [5] for determining the SVD of a matrix **A**. We then calculated the mapping matrix **W** using equation (6) above.

The purpose of calculating the mapping matrix **W** is that it can be multiplied by an unknown vector **a** to get a corresponding vector **b** where **a** is a column vector indicating words in an input text phrase and **b** is an output column vector indicating the relevance of SNOMED concepts. The vector **a** is similar to a column in matrix **A** from the training set. It consists of 1s and 0s that indicate which words are in the text. Vector **b** is similar to a column in matrix **B** from the training set. However, it does not consist only of 1s an 0s. Instead, it consists of values between 0 and 1 where each calculated value gives an indication of how relevant the corresponding SNOMED concept is to the input text phrase. The closer a value is to 1, the more relevant the SNOMED concept is, and the closer a value is to 0, the less relevant the SNOMED concept is.

In order for the user to decide whether a SNOMED concept is relevant to the input text or not given the calculated value between 0 and 1, a threshold needs to be specified. For a given threshold, all SNOMED concepts whose values are greater than or equal to the threshold are said to be relevant, and all SNOMED concepts whose values are less than the threshold are said to be irrelevant to the input text.

The system is implemented in Macintosh Common Lisp (MCL). The calculation of matrix **W** is computationally intensive and only needs to be performed once; the SVD calculation was done on an HP720 and took two CPU minutes.

EVALUATION

To evaluate our methods, we ran the program with 116 sentences taken from hospital admission history and physical examinations for five patients admitted with congestive heart failure. Two of the patients were hospitalized at Stanford University Hospital, and three were hospitalized at Palo Alto Veterans Administration Hospital.

The unit chosen for a single text input was a sentence. In the five history and physical examinations, there were a total of 116 sentences. We ran the program on each of the 116 sentences and compared the SNOMED codes selected by the program with those selected by a human encoder, who was used as a de facto gold standard.

For each of the 116 sentences, the program was run with varying values of threshold. The threshold range was from .1 to .9 in increments of .1. Each SNOMED code selected or not selected at a given threshold for a given sentence was determined to be a true positive, a false positive, a true negative, or a false negative. From these data, sensitivities and specificities were determined. This was done in two ways: One method was to calculate sensitivity (sensitivity = TP / (TP + FN)) and specificity (specificity = TN / (TN + FP)) for each sentence at a given threshold and then average the values over all the sentences. This resulted in an average sensitivity and average specificity for each threshold. The other technique was to count the total true positives and false negatives for all the sentences at a given threshold and calculate an overall sensitivity. Similarly, the true negatives and false positives for all the sentences at a given threshold were counted and an overall specificity was calculated .

The sensitivity and specificity determined for each threshold provided data for plotting a receiver operating characteristic (ROC) curve.

RESULTS

The program successfully extracted SNOMED concepts from input text data, with greater sensitivity at lower thresholds and greater specificity at higher thresholds. The two methods for calculating sensitivities and specificities gave very similar results.

Sample output at varying thresholds for an input sentence taken from a patient history is shown in Fig. 1. Also shown are the number of true positives, false positives, true negatives, and false negatives associated with each threshold.

Test input sentence:
"He describes this as a knot developing in his chest that was constant along with shortness of breath."

Output for threshold .2:

"Chest pain, NOS"	"F-37000"
"Dyspnea, NOS"	"F-20040"
"Hydrochlorothiazide"	"C-72260"
"Lasix Tablets"	"C-C1C6E"
"Lower extremity, NOS"	"T-D9000"
"Mild"	"G-A001"
"Negative for"	"G-A201"

TP=2 FP=5 TN=132 FN=0

Output for threshold .5:

"Chest pain, NOS"	"F-37000"
"Dyspnea, NOS"	"F-20040"

TP=2 FP=0 TN=137 FN=0

Output for threshold .8:

"Dyspnea, NOS"	"F-20040"

TP=1 FP=0 TN=137 FN=1

Fig. 1 Sample Output

In this sample, the expert determined that the correct responses were "Chest pain, NOS" and "Dyspnea, NOS." Each SNOMED concept that was output by the program was either a true positive or a false positive. There were a total of 139 SNOMED concepts known to the program. Every SNOMED concept that was correctly excluded in the output of the program was a true negative, and every SNOMED concept that was missed by the program was a false negative.

The results from the entire data set are shown in Fig. 2, showing performance as a function of threshold. TPR signifies true positive rate (sensitivity) and FPR signifies false positive rate (1 - specificity). The "AVG" data refers to the first method of calculating average sensitivities and specificities. The "TOTAL" data refers to the second method of determining the total number of true positives, false positives, true negatives, and false negatives, and calculating overall sensitivities and specificities from the totals. The ROC curve for the average data is shown in Fig. 3.

Threshold	AVG TPR	AVG FPR
0.1	0.8932	0.0732
0.2	0.8824	0.0336
0.3	0.8457	0.0175
0.4	0.7971	0.0098
0.5	0.7633	0.0055
0.6	0.6830	0.0028
0.7	0.6460	0.0020
0.8	0.5567	0.0011
0.9	0.4657	0.0013

Threshold	TOTAL TPR	TOTAL FPR
0.1	0.9094	0.0730
0.2	0.8960	0.0336
0.3	0.8418	0.0175
0.4	0.8040	0.0097
0.5	0.7703	0.0054
0.6	0.6993	0.0028
0.7	0.6463	0.0020
0.8	0.5495	0.0011
0.9	0.4261	0.0013

Fig. 2
Performance as a Function of Threshold

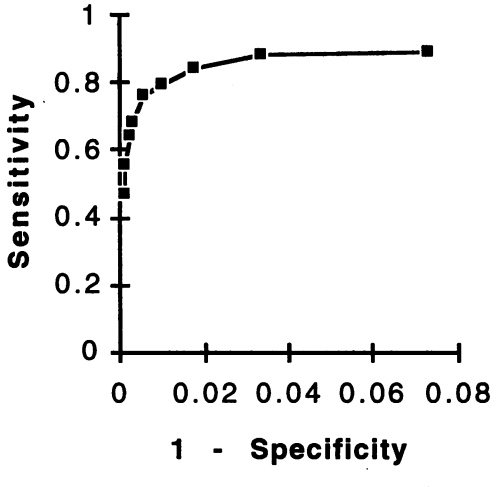

Fig. 3 ROC Curve (Average Data)

DISCUSSION

This program demonstrates that a linear least squares fit approach can be used to automatically assign SNOMED codes to arbitrary natural language text when appropriate training set data have been used to create the mapping function. Yang and Chute [1, 2] demonstrated that the LLSF approach was successful for assigning ICD-9-CM codes to natural language input. Our work differs in that we were attempting to map as many SNOMED codes as were relevant to the entire text of a history and physical examination. In addition, they used a cosine measure to assess similarity between an output result and an ICD-9-CM code, whereas we implemented a threshold model.

The program was trained to extract SNOMED codes from natural language text in history and physical exams for patients with congestive heart failure. When tested, the program was able to assign SNOMED codes for sentences that were relevant to congestive heart failure and that contained concepts similar to those found in the training set. There may be some situations in which a high sensitivity is more important than a high specificity or vice versa. The preferred threshold then would depend on the requirements of the application.

We are now testing several ways in which we could modify the methods to improve performance. For example, we might store all sentences or clauses from the documents in the columns of matrix A rather than phrases that were selected by a physician as medically relevant. Similarly, we could keep our test documents the same but change the method of running the program on test documents by making each unknown input a paragraph rather than a sentence.

The evaluation performed for this study assessed the ability of the program to meet the performance level set by the physician who encoded the training set and provided the gold standard codes for the evaluation. Thus, it evaluated the validity of the method for reproducing the performance of a single encoder. It did not evaluate the degree to which that encoder was a valid gold standard.

Further evaluation should assess the ability of the program to meet the expectations of a group of physicians who have their own particular biases and no prior information about the program. This would be a measure of the knowledge stored in the system as well as an evaluation of the method. In addition, further work could focus on expanding the domain to include patients with other conditions besides congestive heart failure. One might create separate mapping matrices for different diagnoses and patient complaints. A patient's medical record document could then be processed by applying the appropriate matrices, in sequence, for each of the patient's diagnoses or major complaints. Another option would be to store training data on multiple domains all in the same matrix, but scaling may be a problem as the matrices expand greatly in size and computations become more complex.

In conclusion, the LLSF and SVD approach may be a useful technique for automatic extraction of standardized vocabulary concepts from medical record text documents. This study suggests that it is useful for a small domain. Further work using larger training sets would be required to determine the utility of this approach in larger domains.

ACKNOWLEDGEMENTS

Computing facilities were provided for this work by the Center for Advanced Medical Informatics at Stanford, which is supported by NLM grant LM05305. Dr. Oliver is supported by AHCPR training grant HS 00028. Dr. Altman is supported by NIH grant LM05652 and the Culpeper Foundation.

REFERENCES

1. Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L. The Systematized Nomenclature of Human and Veterinary Medicine, SNOMED International, College of American Pathologists, Northfield, IL, 1993.

2. Yang Y, Chute CG. An application of least squares fit mapping to clinical classification. Proceedings of the 16th Annual Sumposium on Computer Applications in Medical Care, 1992, pp. 460-464.

3. Yang Y, Chute CG. A linear least squares fit mapping method for information retrieval from natural language texts. Proceedings of the 14th International Conference on computational Linguistics-92, Nantes, Aug. 23-28, 1992, 447-453.

4. International Classification of Diseases, 9th Revision, Clinical Modification, Fourth Edition. Practice Management Information Corporation, Los Angeles, CA, 1993.

5. Forsythe GE, Malcolm MA, Moler CG. Computer Methods for Mathematical Computations. Prentice-Hall, Inc. Englewood Cliffs, NJ, 1977.